09/807648 PCT Rec'd 16 APR 2001
PCT/GB99/03322

Speech Processing

This invention relates to apparatus and a method for estimating the speech level of a speaker exposed to an environment containing a variable amount of acoustic noise. In particular, but not exclusively, the invention relates to such apparatus and methods for use in speech recognition.

The central process in automatic speech recognition is the comparison between some representation of the speech to be recognised and a set of reference models corresponding to speech sounds or words or other units. It is important that the level of the speech signal represented in the recogniser should be close to that expected by the models.

Because speech sounds vary in their intrinsic loudness, measuring overall speech level is not a trivial process. It is necessary either to take a large enough sample of the speech that the variations occurring between speech sounds average out, or to compare an utterance whose level is to be measured with an utterance at some known level whose phonetic content is the same. In this second method, phonetically identical speech sounds can be compared, but it does require a knowledge of the content of the utterance to be measured.

We have realised that it is in fact possible to estimate variations in the likely level of the speech signal in acoustically noisy environments by measuring the ambient noise level and using a phenomenon known as the Lombard

Effect to determine the likely speech levels. The Lombard Effect is the phenomenon that when people are exposed to noise their speech changes and becomes generally becomes louder. If no adjustment is made for the Lombard Effect in an automatic speech recognition system there will be a mismatch between the level of the speech to be recognised and the expected level. In principle, this could be corrected by observing the speech level and adjusting the gain of an amplifier in the recogniser to compensate for the variation in level. However, in some circumstances this is not a practical arrangement. For example, in a car the noise level can change from one utterance to another following changes in the speed of the car or in the road surface, or because a window is wound down. A gain setting based on the previous utterance will then be inappropriate. In some circumstances, it might be possible to wait until the utterance was complete, measure the speaking level, adjust the recorded utterance to normalise this level, and only then submit it to the recogniser. However, this process would introduce a delay in the response of the recogniser, which for many applications would be unacceptable.

In one aspect, this invention provides apparatus for predicting the speech level of a speaker exposed to an environment containing a variable level of ambient acoustic noise, the apparatus comprising means for measuring said ambient acoustic noise level, and processing means for using said measured acoustic noise level to predict the likely

speech level.

In this apparatus, as the noise level in the environment in which the speaker is located changes between utterances, so his speech level is likely to rise and fall in accordance with the Lombard Effect, and the apparatus predicts the likely speech level. We have found that the likely speech level can be predicted with reasonable accuracy by measuring the noise immediately adjacent to an utterance; measuring the level of a steady noise is quite simple and can be carried out with just a short sample of the noise. The apparatus preferably also uses a measure of the speech level and the corresponding noise level relating to a previous or standardised utterance.

The ambient acoustic noise level could be measured before, after or even during utterance of a word or phrase, and it is preferred for the measurement to be made close in time to the utterance to reduce the possibility of the prediction of the likely speech level being inaccurate due to a significant shift in noise level between measurement and the actual utterance.

It is preferred for the measuring means to measure the ambient acoustic noise level immediately before the utterance, the estimate of speech level being determined before or as the utterance is made rather than thereafter. Alternatively the measurement may be after the utterance.

The apparatus preferably includes means operable to define, for each utterance, an utterance period comprising a first time period for measuring said acoustic noise level

and a second time period during which said utterance is made.

Thus in a preferred embodiment, the apparatus includes a user input device (such as e.g. a switch) and a timer and control means for defining said first noise measuring period, and said second speech measuring and/or recording period, the end of said first period being indicated to said user.

In a particularly preferred aspect, said apparatus is responsive to a succession of one or more utterances by a speaker and said measuring means measures the ambient noise level prevailing at each of said utterances to provide a series of noise measurements and said apparatus includes means for measuring the speech level of an utterance, and said processing means uses at least two of said noise measurements, together with the measurement of the speech level of the immediately previous utterance, to produce the prediction of the speech level of the most recent utterance.

In one example, where the noise is measured immediately before an utterance, the processing apparatus means predicts the speech level S_1^* of an utterance (1) on the basis of the following expression:

$$S_1^* = S_o+f(N_o-N_1)$$

where S_o is the speech level of the immediately previous utterance; N_1,N_o are the noise levels prevailing immediately before the utterance whose speech level is to be estimated, and immediately before the next previous utterance respectively, and f(x) is a function relating changes in the

noise level in which the speaker is situated to the speaker's speech level.

The function is preferably monotonic increasing, and in a simple case is a multiplying factor less than 1. The multiplying factor may typically be a positive value in the range of from 0 to 0.6, and in one example is 0.32.

Alternatively the function may be a more complex function of the noise level difference. Likewise, the function may be modified to take account of more than just two noise level measurements; thus information relating to the speech levels of several previous utterances, together with the associated noise levels may be aggregated to predict the speech level of the next utterance.

In another aspect, this invention provides speech recognition or processing apparatus including predicting apparatus as set out above for use in adjusting the gain of the speech signal prior to recognition processing.

In yet another aspect, this invention provides a method for predicting the speech level of a speaker exposed to an environment containing a variable level of ambient acoustic noise, said method comprising the steps of:-

measuring said ambient acoustic noise level, and

processing said measured acoustic noise level to produce a prediction of the likely speech level.

In a further aspect, this invention provides a method for controlling the gain in a speech recognition or processing system, which comprises controlling the gain of the speech signal in accordance with a prediction of the

speech level obtained by the above method.

Whilst the invention has been described above, it extends to any inventive combination of the features set out above or in the following descriptions.

The invention may be performed in various ways, and an embodiment thereof will now be described by way of example only, reference being made to the accompanying drawing in which:-

Figure 1 is a block diagram of a speech recogniser incorporating speech level prediction in accordance with the invention.

The illustrated embodiment implements a system which applies knowledge of variation in the ambient acoustic noise level and its likely effect on the speech level to predict the speech level in the next utterance to be recognised by a speech recogniser. It is assumed that the variation in noise level over the duration of a single utterance is small compared with the variations occurring between utterances, and also that the noise has sufficient short-term stationarity that its level can be measured from a brief sample.

Referring to Figure 1, the speech recognition system comprises a microphone 10 whose output is subjected to voice processing at 12 before analogue to digital conversion at 14. The digital signal passes via a digital gain device 16 to a processor 18 which incorporates a recogniser 20 and a speech level estimator 22. The speech recogniser may be of any suitable type and examples of suitable recognisers will

be well known to those skilled in the art. The processor 18 also receives an input from a switch 24 acting as a user input device, and can issue warning tones to the user through a sounder 26.

The system illustrated is intended for use in a noisy environment whose noise level varies. In use, the user alerts the system when he wants to make an utterance to be recognised, by closing the switch 24. The processor then defines an utterance frame, comprising a first short time period, during which the ambient noise is sampled, followed by issuing a tone on the sounder 26, which indicates to the user that he may speak, followed by a second period during which the speech signal is sampled and sent to the recogniser 20. The second period is longer than the first period and sufficiently long to contain the longest utterance to be recognised. There are a number of ways of delimiting the second period other than providing a period of set duration. For example the length of the period may be user designated, e.g. by the user keeping the button pressed or pressing the button again. Alternatively, the processor may listen for a period of silence, or it may infer the end of a command based on an analysis of the grammar of the utterance. In addition, instead of using a switch, the start of the utterance frame may be marked by the user uttering a codeword.

Since it is known that speech levels vary with noise level, it is possible to predict a change in the speech level in an utterance from a change in the noise level. The

speech and noise levels, S_0 and N_0, (in dB units) are measured by the processor in one noise condition. The new noise level, N_1, in the first period of the next utterance, just before the start of an utterance to be recognised, is also measured by the processor. The difference in the two noise levels, N_0-N_1, is then determined and used by the processor, together with knowledge of the speech level, S_0 of the previous utterance, to predict the speech level, S_1, of the new utterance. We can write $S^*_1 = S_0 + f(N_0-N_1)$, where S^*_1 is a prediction estimate of S_1 and $f(x)$ is the function relating changes in the noise level in the speaker's ears to the speaker's speech level. In the simplest arrangement, the function is a multiplying factor less than 1, but it can also be a more complex function of the noise level difference. In practice we have determined empirically that the speech level good results are achieved in one application by using a multiplying factor of typically 0.3 although positive values between 0 and 0.6 should all provide some improvement. It may be assumed to be the same for all speakers or may be estimated separately for each speaker.

Since the measurements of the reference speech and noise levels, S_0 and N_0, respectively, are subject to measurement errors, it may be preferred to aggregate the information contributing to the prediction of S_1 from several previous utterances and noise estimates. The computation of S^*_1 described in the previous paragraph can be replaced by an average over several previous utterances. This may be a

simple average or it may be a weighted average, the weights possibly depending on factors such as the time difference between the various reference utterances and S_1 and on the relative durations of the various reference utterances. For example the computation may take account of any time effects. For example it may be found that, when exposed to a particular level of ambient noise that the speaker's speech level rises over an initial period and then decreases, in a temporal filtering effect.

Having determined an estimate of the speech level of the new utterance, the processor controls the gain of the signal accordingly. The gain may be adjusted at various points; it may be adjusted whilst the signal is still in the analogue domain or it may be achieved by digital scaling as shown by the digital gain device 16. A further alternative is to manipulate the fast fourier transform (FFT) values in the speech recogniser. If a cepstrum is computed, the signal may be scaled by adding an appropriate constant to the C_o coefficient. In a further arrangement, the system may compensate for increases or decreases in the speech level by adjusting the effective speech levels that the models in the recogniser represent.

The gain may take into account factors other than simply the level of the background noise; for example it could also take account of its spectral structure.

The output of the recogniser may be used in any convenient form. For example it could be used to enable a person to issue spoken commands to equipment.